
Appendix A: Technical Notes

Information on the technical aspects of TIMSS 2003 is provided below. More detailed information can be found in the TIMSS 2003 Technical Report (Martin, Mullis, and Chrostowski 2004).

Data Collection

The TIMSS 2003 data were collected by each country, following international guidelines and specifications. TIMSS required that countries select random, nationally representative samples of schools and students. TIMSS countries were asked to identify eligible students based on a common set of criteria, allowing for adaptation to country-specific situations. In IEA studies such as TIMSS, the target population for all countries is called the international desired population. For the fourth-grade assessment, the international desired population consisted of all students in the country who were enrolled in the upper of the two adjacent grades that contained the greatest proportion of 9-year-olds at the time of testing. In the United States and most other countries, this corresponded to fourth grade. For the eighth-grade assessment, the international desired population consisted of all students in the country who were enrolled in the upper of the two adjacent grades that contained the greatest proportion of 13-year-olds at the time of testing. In the United States and most other countries, this corresponded to eighth grade.

TIMSS used a two-stage stratified cluster sampling design. The first stage made use of a systematic probability-proportionate-to-size (PPS) technique to select schools. Although countries participating in TIMSS were strongly encouraged to secure the participation of schools selected in the first stage, it was anticipated that a 100 percent participation rate for schools would not be possible in all countries. Therefore, two replacement schools were identified for each originally sampled school, a priori. As each school was selected, the next school in the sampling frame was designated as a replacement school should the originally sampled school choose not to participate in the study. Should the originally sampled school and the replacement school choose not to participate, a second replacement school was chosen by going to the next school in the sampling frame.

The second stage consisted of selecting classrooms within sampled schools. At the classroom level, TIMSS sampled intact mathematics classes that were offered to students in the target grades. In most countries, one mathematics classroom per school was sampled, although some countries, such as the United States, chose to sample two mathematics classrooms per school.

Exclusions in the TIMSS Sample

All countries were required to define their national desired population to correspond as closely as possible to the definition of the international desired population. In some cases, countries needed to exclude schools and students in remote geographical locations or to exclude a segment of the education system. Any exclusions from the international desired population were clearly documented. Countries were expected to keep the excluded population to no more than 10 percent of the national desired population. Exclusions could take place at the school level, within schools, or both. Participants could exclude schools from the sampling frame for the following reasons:

- Locations were geographically remote;
- Size was extremely small;
- Curriculum or school structure was different from the mainstream education system; or
- Instruction provided was only to students in the categories defined as “within-school exclusions.”

Within schools, exclusion decisions were limited to students who, because of some disability, were unable to take part in the TIMSS assessment. The general TIMSS rules for defining within-school exclusion included the following three groups:

- *Intellectually disabled students.* These students were considered, in the professional opinion of the school principal or other qualified staff members, to be intellectually disabled, or had been so diagnosed in psychological tests. This category included students who were emotionally or mentally unable to follow even the general instructions of the TIMSS test. It did not include students who merely exhibited poor academic performance or discipline problems.
- *Functionally disabled students.* These students were permanently physically disabled in such a way that they could not participate in the TIMSS assessment. Functionally disabled students who could perform were included in the testing.

- *Non-native-language speakers.* These students could not read or speak the language of the assessment and so could not overcome the language barrier of testing. Typically, a student who had received less than 1 year of instruction in the language of the assessment was excluded, but this definition was adapted in different countries.

School-level and within-school exclusion rates for TIMSS 2003 are detailed in the next section. Exclusion rates for TIMSS 1995 can be found in chapter 2 of Martin and Kelly (1997); exclusion rates for TIMSS 1999 can be found in appendix 2 of Gonzales et al. (2000).

Response Rates

Based on the sample of schools and students that participated in the assessment, countries were assigned to one of four following categories:

Category 1: met requirements

- An unweighted or weighted school response rate without replacement of at least 85 percent and an unweighted or weighted student response rate of at least 85 percent

or

- The product of the weighted school response rate without replacement and the weighted student response rate of at least 75 percent.

Category 2: met requirements after replacements

- If the requirements for category 1 are not met but the country had either an unweighted or weighted school response rate without replacement of at least 50 percent and had either
- An unweighted or weighted school response rate with replacement of at least 85 percent and a weighted student response rate of at least 85 percent

or

- The product of the weighted school response rate with replacement and the weighted student response rate of at least 75 percent.

Category 3: close to meeting requirements after replacements

- If the requirements for category 1 or 2 are not met but the country had either an unweighted or weighted school response rate without replacement of at least 50 percent and
- The product of the weighted school response rate with replacement and the weighted student response rate near 75 percent.

Category 4: failed to meet requirements

- Unacceptable sampling response rate even when replacement schools are included.

In this report, countries in category 1 appear in the tables and figures without annotation; countries in category 2 are annotated in the tables and figures; countries in category 3 are enclosed with parentheses in the tables and figures, as is the case, for example, of the United States and Morocco at eighth grade. Finally, countries in category 4 are not shown in tables or figures in this report. In addition, annotations are included when the exclusion rate exceeds 10 percent. Latvia is designated as Latvia-LSS (Latvian-speaking schools) in some analyses because data collection in 1995 and 1999 was limited to only those schools in which instruction was in Latvian. Finally, Belgium is annotated as Belgium-Flemish because only the Flemish education system in Belgium participated in TIMSS.

Information on the populations assessed and participation rates is provided in table A1. Details on the number of TIMSS participating schools and students in each of the participating countries are provided in table A2.

Table A1. Coverage of TIMSS grade 4 and 8 target population and participation rates, by country: 2003

Country	Grade 4						Combined weighted school and student participation rate
	Years of formal schooling	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before replacement	Weighted school participation rate after replacement	Weighted student participation rate	
Armenia	4	100	3	99	99	91	90
Australia	4 or 5	100	3	78	90	94	85
Belgium-Flemish	4	100	6	89	99	98	97
Chinese Taipei	4	100	3	100	100	99	99
Cyprus	4	100	3	100	100	97	97
England	5	100	2	54	82	93	76
Hong Kong SAR ¹	4	100	4	77	88	95	83
Hungary	4	100	8	98	99	94	93
Iran, Islamic Republic of	4	100	6	100	100	98	98
Italy	4	100	4	97	100	97	97
Japan	4	100	1	100	100	97	97
Latvia	4	100	4	91	94	94	88
Lithuania	4	92	5	92	96	92	87
Moldova, Republic of	4	100	4	97	100	97	97
Morocco	4	100	2	87	87	93	81
Netherlands	4	100	5	52	87	96	84
New Zealand	4.5 - 5.5	100	4	87	98	95	93
Norway ²	4	100	4	89	93	95	88
Philippines	4	100	5	78	85	95	81
Russian Federation	3 or 4	100	7	99	100	97	97
Scotland	5	100	1	64	83	92	77
Singapore	4	100	0	100	100	98	98
Slovenia	3 or 4	100	1	95	99	92	91
Tunisia	4	100	1	100	100	99	99
United States	4	100	5	70	82	95	78

See notes at end of table.

Table A1. Coverage of TIMSS grade 4 and 8 target population and participation rates, by country: 2003—Continued

Country	Grade 8						
	Years of formal schooling	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before replacement	Weighted school participation rate after replacement	Weighted student participation rate	Combined weighted school and student participation rate
Armenia	8	100	3	99	99	90	89
Australia	8 or 9	100	1	81	90	93	83
Bahrain	8	100	0	100	100	98	98
Belgium-Flemish	8	100	3	82	99	97	94
Botswana	8	100	3	98	98	98	96
Bulgaria	8	100	0	97	97	96	92
Chile	8	100	2	98	100	99	99
Chinese Taipei	8	100	5	100	100	99	99
Cyprus	8	100	3	100	100	96	96
Egypt	8	100	3	99	100	97	97
Estonia	8	100	3	99	99	96	95
Ghana	8	100	1	100	100	93	93
Hong Kong SAR ¹	8	100	3	74	83	97	80
Hungary	8	100	9	98	99	95	94
Indonesia	8	80	0	98	100	99	99
Iran, Islamic Republic of	8	100	6	100	100	98	98
Israel	8	100	23	98	99	95	94
Italy	8	100	4	96	100	97	97
Japan	8	100	1	97	97	96	93
Jordan	8	100	1	100	100	96	96
Korea, Republic of	8	100	5	99	99	99	98
Latvia	8	100	4	92	94	89	83
Lebanon	8	100	1	93	95	96	91
Lithuania	8	89	3	92	95	89	84

See notes at end of table.

Table A1. Coverage of TIMSS grade 4 and 8 target population and participation rates, by country: 2003—Continued

Country	Grade 8						
	Years of formal schooling	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before replacement	Weighted school participation rate after replacement	Weighted student participation rate	Combined weighted school and student participation rate
Macedonia, Republic of	8	100	12	94	99	97	96
Malaysia	8	100	4	100	100	98	98
Moldova, Republic of	8	100	1	99	100	96	96
Morocco	8	69	1	79	79	91	71
Netherlands	8	100	3	79	87	94	81
New Zealand	8.5 - 9.5	100	4	86	97	93	90
Norway	7	100	2	92	92	92	85
Palestinian National Authority	8	100	0	100	100	99	99
Philippines	8	100	1	81	86	96	82
Romania	8	100	1	99	99	98	98
Russian Federation	7 or 8	100	6	99	99	97	96
Saudi Arabia	8	100	1	95	97	97	94
Scotland	9	100	0	76	85	89	76
Serbia	8	81	3	99	99	96	96
Singapore	8	100	0	100	100	97	97
Slovak Republic	8	100	5	96	100	95	95
Slovenia	7 or 8	100	1	94	99	93	91
South Africa	8	100	1	89	96	92	88
Sweden	8	100	3	97	99	89	87
Tunisia	8	100	2	100	100	98	98
United States	8	100	5	71	78	94	73

¹Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.

²Norway Grade 4: 4 years of formal schooling, but first grade is called "first grade/preschool."

NOTE: Only countries that completed the necessary steps for their data to appear in the reports from the International Study Center are listed. In addition to the countries listed above, four separate jurisdictions participated in TIMSS 2003: the provinces of Ontario and Quebec in Canada; the Basque region of Spain; and the state of Indiana. Yemen participated in TIMSS 2003 but due to difficulties with the data, does not appear in this report. England participated in TIMSS 2003 but did not meet the minimum sampling requirements at grade 8. Information on these jurisdictions can be found in the international *TIMSS 2003 Technical report* (Martin, Mullis, and Chrostowski 2004). SOURCE: Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., and Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report: Findings from the IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

Table A2. TIMSS grade 4 and 8 student and school samples, by country: 2003

Country	Grade 4						
	Schools in original sample	Eligible schools in sample	Schools in original sample that participated	Replacement schools	Total schools that participated	Sampled students in participating schools	Students assessed
Armenia	150	150	148	0	148	6,275	5,674
Australia	230	227	178	26	204	4,675	4,321
Belgium-Flemish	150	150	133	16	149	4,866	4,712
Chinese Taipei	150	150	150	0	150	4,793	4,661
Cyprus	150	150	150	0	150	4,536	4,328
England	150	150	79	44	123	3,917	3,585
Hong Kong SAR ¹	150	150	116	16	132	4,901	4,608
Hungary	160	159	156	1	157	3,603	3,319
Iran, Islamic Republic of	176	171	171	0	171	4,587	4,352
Italy	172	171	165	6	171	4,641	4,282
Japan	150	150	150	0	150	4,690	4,535
Latvia	150	149	137	3	140	3,980	3,687
Lithuania	160	160	147	6	153	5,701	4,422
Moldova, Republic of	153	151	147	4	151	4,162	3,981
Morocco	227	225	197	0	197	4,546	4,264
Netherlands	150	149	77	53	130	3,080	2,937
New Zealand	228	228	194	26	220	4,785	4,308
Norway	150	150	134	5	139	4,706	4,342
Philippines	160	160	122	13	135	5,225	4,572
Russian Federation	206	205	204	1	205	4,229	3,963
Scotland	150	150	94	31	125	4,283	3,936
Singapore	182	182	182	0	182	6,851	6,668
Slovenia	177	177	169	5	174	3,410	3,126
Tunisia	150	150	150	0	150	4,408	4,334
United States	310	300	212	36	248	10,795	9,829

See notes at end of table.

Table A2. TIMSS grade 4 and 8 student and school samples, by country: 2003—Continued

Country	Grade 8						
	Schools in original sample	Eligible schools in sample	Schools in original sample that participated	Replacement schools	Total schools that participated	Sampled students in participating schools	Students assessed
Armenia	150	150	149	0	149	6,388	5,726
Australia	230	226	186	21	207	5,286	4,791
Bahrain	67	67	67	0	67	4,351	4,199
Belgium-Flemish	150	150	122	26	148	5,161	4,970
Botswana	152	150	146	0	146	5,388	5,150
Bulgaria	170	169	163	1	164	4,489	4,117
Chile	195	195	191	4	195	6,528	6,377
Chinese Taipei	150	150	150	0	150	5,525	5,379
Cyprus	59	59	59	0	59	4,314	4,002
Egypt	217	217	215	2	217	7,259	7,095
Estonia	154	152	151	0	151	4,242	4,040
Ghana	150	150	150	0	150	5,690	5,100
Hong Kong SAR ¹	150	150	112	13	125	5,204	4,972
Hungary	160	157	154	1	155	3,506	3,302
Indonesia	150	150	148	2	150	5,884	5,762
Iran, Islamic Republic of	188	181	181	0	181	5,215	4,942
Israel	150	147	143	3	146	4,880	4,318
Italy	172	171	164	7	171	4,628	4,278
Japan	150	150	146	0	146	5,121	4,856
Jordan	150	140	140	0	140	4,871	4,489
Korea, Republic of	151	150	149	0	149	5,451	5,309
Latvia	150	149	137	3	140	4,146	3,630
Lebanon	160	160	148	4	152	4,030	3,814
Lithuania	150	150	137	6	143	6,619	4,964

See notes at end of table.

Table A2. TIMSS grade 4 and 8 student and school samples, by country: 2003—Continued

Country	Grade 8						
	Schools in original sample	Eligible schools in sample	Schools in original sample that participated	Replacement schools	Total schools that participated	Sampled students in participating schools	Students assessed
Macedonia, Republic of	150	150	142	7	149	4,028	3,893
Malaysia	150	150	150	0	150	5,464	5,314
Moldova, Republic of	150	149	147	2	149	4,262	4,033
Morocco	227	165	131	0	131	3,243	2,943
Netherlands	150	150	118	12	130	3,283	3,065
New Zealand	175	174	149	20	169	4,343	3,801
Norway	150	150	138	0	138	4,569	4,133
Palestinian National Authority	150	145	145	0	145	5,543	5,357
Philippines	160	160	132	5	137	7,498	6,917
Romania	150	149	148	0	148	4,249	4,104
Russian Federation	216	216	214	0	214	4,926	4,667
Saudi Arabia	160	160	154	1	155	4,553	4,295
Scotland	150	150	115	13	128	3,962	3,516
Serbia	150	150	149	0	149	4,514	4,296
Singapore	164	164	164	0	164	6,236	6,018
Slovak Republic	180	179	170	9	179	4,428	4,215
Slovenia	177	177	169	5	174	3,883	3,578
South Africa	265	265	241	14	255	9,905	8,952
Sweden	160	160	155	4	159	4,941	4,256
Tunisia	150	150	150	0	150	5,106	4,931
United States	301	296	211	21	232	9,891	8,912

¹Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.

NOTE: Only countries that completed the necessary steps for their data to appear in the reports from the International Study Center are listed. In addition to the countries listed above, four separate jurisdictions participated in TIMSS 2003: the provinces of Ontario and Quebec in Canada; the Basque region of Spain; and the state of Indiana. Yemen participated in TIMSS 2003 but due to difficulties with the data, does not appear in this report. England participated in TIMSS 2003 but did not meet the minimum sampling requirements at grade 8. Information on these jurisdictions can be found in the international *TIMSS 2003 Technical report* (Martin, Mullis, and Chrostowski 2004). SOURCE: Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., and Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report: Findings from the IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

Sampling, Data Collection, and Response Rates in the United States

The TIMSS 2003 school sample was drawn for the United States in November 2002. The sample design for this school sample was developed to follow international requirements as given in the TIMSS sampling manual. The U.S. sample for 2003 was a two-stage sampling process with the first stage a sample of schools, and the second stage a sample of students' classrooms from the target grade in sampled schools. Unlike TIMSS 1995 and 1999, the sample was not clustered at the geographic level for TIMSS 2003.

This change was made in an effort to reduce the design effects and to spread the respondent burden across schools districts as much as possible.

The sample design for TIMSS was a stratified systematic sample, with sampling probabilities proportional to measures of size. The U.S. TIMSS fourth-grade sample had two explicit strata based on poverty. A high poverty school was defined as one in which 50 percent or more of the students were eligible for participation in the federal free or reduced-price lunch program; high poverty schools were oversampled (Ferraro and Rust 2003) This variable

was not available for private schools, so they were all treated as low poverty schools. The target sample sizes were 120 high-poverty and 190 low-poverty schools.

Within the poverty strata, there are four categorical implicit stratification variables: type of school (public or private), region of the country¹⁹ (Northeast, Southeast, Central, West), type of location relative to populous areas (eight levels), minority status (above or below 15 percent). The last sort key within the implicit stratification was by grade enrollment in descending order.

The TIMSS eighth-grade sample had no explicit stratification. The frame was implicitly stratified (i.e., sorted for sampling) by four categorical stratification variables: type of school (public or private), region of the country, type of location relative to populous areas (eight levels), minority status (above or below 15 percent). The last sort key within the implicit stratification was by grade enrollment in descending order.

At the same time that the TIMSS sample was selected, replacement schools were identified following the TIMSS guidelines by assigning the two schools neighboring the sampled school on the frame as replacements. There were several constraints on the assignment of substitutes. One sampled school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school. Furthermore, substitutes were required to be in the same implicit stratum as the sampled school. If the sampled school was the first or last school in the stratum, then the second school following or preceding the sampled school was identified as the substitute. One was designated a first replacement and the other a second replacement. If an original school refused to participate, the first replacement was then contacted. If that school also refused to participate, the second school was then contacted.

The schools were selected with probability proportionate to the school's estimated enrollment of fourth- and eighth-grade students from the 2003 NAEP school frame with 2000-01 school data. The data for public schools were from the Common Core of Data (CCD), and the data for private schools was from the Private School Survey (PSS). Any school containing a fourth or an eighth grade as of the school year 2000-01 was included on the school

sampling frame. Participating schools provided lists of fourth- or eighth-grade classrooms, and one or two intact mathematics classrooms were selected within each school in an equal probability sample. The overall sample design for the United States was intended to approximate a self-weighting sample of students as much as possible, with each fourth- or eighth-grade student having an equal probability of being selected.

The U.S. TIMSS fourth-grade school sample consisted of 310 schools, of which 300 were eligible schools and 212 agreed to participate. The school response rate before replacement was 70 percent (weighted; 71 percent unweighted). The weighted school response rate before replacement is given by the formula:

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i},$$

where Y denotes the set of responding original sample schools with age-eligible students, N denotes the set of eligible non-responding original sample schools, W_i denotes the base weight for school i, $W_i = 1/P_i$, where P_i denotes the school selection probability for school i, and E_i denotes the enrollment size of age-eligible students, as indicated on the sampling frame.

In addition to the 212 participating schools, 36 replacement schools also participated for a total of 248 participating schools at the fourth grade in the United States.

A total of 10,795 students were sampled for the fourth-grade assessment. Of these students, 49 were withdrawn from the school before the assessment was administered. Of the eligible 10,746 sampled students, an additional 429 students were excluded using the criteria described above, for a weighted exclusion rate of 5 percent. Of the 10,317 remaining sample students, a total of 9,829 students participated in the assessment in the United States, since 488 students were absent. The student participation rate was 95 percent.

The combined school and students weighted and unweighted response rate of 78 percent after replacement schools were included was achieved (66 percent weighted and 67 percent unweighted

¹⁹Region is the 'state-based' region (NAEPRG_S on the output files). Northeast consists of Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. Central consists of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. West consists of Alaska, Arizona, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Texas, Utah, Washington, Oregon, California, and Wyoming. Southeast consists of Alabama, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.

without replacement). As a result, the U.S. data for fourth-grade students are annotated to indicate that international guidelines for participation rates were met only after replacement schools were included.

The U.S. TIMSS eighth-grade school sample consisted of 301 schools, of which 296 were eligible schools and 211 agreed to participate. The school response rate before replacement was 71 percent (weighted and unweighted). In addition to the 211 participating schools, 21 replacement schools also participated for a total of 232 participating schools at the eighth grade in the United States.

A total of 9,891 students were sampled for the assessment. Of these students, 90 were withdrawn from the school before the assessment was administered. Of the eligible 9,801 sampled students, an additional 279 students were excluded using the criteria described above, for a weighted exclusion rate of 5 percent. Of the 9,522 remaining sample students, a total of 8,912 students participated in the assessment in the United States, since 610 students were absent. The student participation rate was 94 percent (weighted and unweighted). The combined school and students weighted and unweighted response rate of 73 percent after replacement schools were included was achieved (66 percent without replacement schools). As a result, the U.S. data for eighth-grade students are in parentheses to indicate that United States did not meet international sampling guidelines.

NCES standards require a nonresponse bias analysis if the school level response rate is below 80 percent (using the base weight). Since the U.S. school response rates at the fourth and eighth grades were below 80 percent, even with replacements, NCES required an analysis of the potential magnitude of nonresponse bias at the school level. To accomplish this analysis, two methods were chosen (Van de Kerckhove and Ferraro forthcoming). The first method was focused exclusively on the original sample of schools, treating all those that were substituted as nonrespondents. A second method focused on the final sample of schools (including replacements), treating as nonrespondents those schools from which a final response was not received. Both methods were used to analyze the U.S. TIMSS fourth- and eighth-grade data for potential bias.

In order to compare TIMSS respondents and nonrespondents it was necessary to match the sample of schools back to the sample frame to detect as many characteristics as possible that might provide information about the presence of nonresponse bias. Comparing characteristics for respondents and nonrespondents is not always a good measure of nonresponse bias if the characteristics are unrelated or weakly related to more substantive items in the survey. However, this is often the only approach available. The characteristics that were analyzed based on the sampling frame were taken from the 2000-2001 Common Core of Data (CCD) for public schools, and from the 2000-2001 Private School Survey (PSS) for private schools. For categorical variables, the distribution of the characteristics for respondents was compared with the distribution for all schools. The hypothesis of independence between a given school characteristic and the response status (whether or not participated) was tested using a Rao-Scott modified Chi-square statistic. For continuous variables, summary means were calculated. The 95 percent confidence interval for the difference between the mean for respondents and the mean for all schools was tested to see whether or not it included zero. In addition to these tests, logistic regression models were set up to identify whether any of the school characteristics were significant in predicting response status because logistic regression allows investigation of all variables at the same time.

Public and private schools were modeled together using the following variables: community type; public/religious affiliation; NAEP region; poverty level; number of students enrolled in fourth or eighth grade; total number of students; percentage Asian or Pacific Islander students; percentage Black, non-Hispanic students; percentage Hispanic students; percentage American Indian or Alaska Native students; and percentage White, non-Hispanic students.

The investigation into nonresponse bias at the school level for TIMSS fourth grade generally showed that there was no statistically significant relationship between response status and the majority of school characteristics available for analysis. For the original sample of schools in TIMSS fourth grade, schools in the Northeast were less likely to respond than schools in the West, Southeast or Central regions of the coun-

try. However, the regression did not confirm this result. The results for the final sample of schools showed a significant effect on the percentage of Black, non-Hispanic students (responding schools had more Black, non-Hispanic students than non-responding schools). However, the regression did not confirm this result.

The investigation into nonresponse bias at the school level for TIMSS eighth grade showed that, for the original sample of schools, responding schools were more likely to be in rural areas than in central city or urban fringe areas, have fewer students than non-responding schools, have fewer Hispanic students, and were more likely to be Catholic or public schools. However, the regression confirmed only that responding schools in the original sample were more likely to be from rural areas and have fewer students than non-responding schools. The number of Hispanic students in responding schools and their public/religious affiliation were not confirmed by the regression. The results with the final sample of schools were more complicated. The total number of students remained significant, but the additional variable of public/religious affiliation also appeared to be significantly related to response rate according to the logistic regression. Public and Catholic schools were more likely to respond than private, non-sectarian and private-other religious schools. Finally, while the first analysis indicated that schools in rural areas were more likely to respond than schools in the central city or urban fringe, this was not confirmed by the logistic regression.

The results of these analyses suggest that there is no statistically significant relationship between response status and the majority of the school characteristics tested, with the exception of the variables noted above at each grade level. The potential for nonresponse bias exists however. It is difficult to assess the amount of any bias in the survey as a result of the associations that exist.

It is also not clear what effect the weighting adjustments for nonresponse have on any bias. In general, these weighting adjustments cannot address all of the potential bias, only some of it. There is no evaluation of how much effect the weighting adjustments have on the bias.

Test Development

TIMSS is a cooperative effort involving representatives from every country participating in the study. For TIMSS 2003, the development effort began with a revision of the frameworks that are used to guide the construction of the assessment (Mullis et al. 2001). The framework was updated to reflect changes in the curriculum and instruction of participating countries. Extensive input from experts in mathematics and science education, assessment, curriculum, and representatives from national educational centers around the world contributed to the final shape of the frameworks. Maintaining the ability to measure change over time was an important factor in revising the frameworks.

As part of the TIMSS dissemination strategy, approximately one-third of the 1995 fourth-grade assessment items and one-half of the 1999 eighth-grade assessment items were released for public use. To replace assessment items that had been released in earlier years, countries submitted items for review by subject-matter specialists, and additional items were written to ensure that the content, as explicated in the frameworks, was covered adequately. Items were reviewed by an international Science and Mathematics Item Review Committee and pilot-tested in most of the participating countries. Results from the field test were used to evaluate item difficulty, how well items discriminated between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias towards or against individual countries or in favor of boys or girls. As a result of this review, 243 of the 435 new fourth-grade items were selected for inclusion in the assessment. In total, there were 313 mathematics and science items included in the fourth-grade TIMSS assessment booklets. At eighth grade, the review of the item statistics from the field test led to the inclusion of 230 of the 386 new eighth-grade items in the assessment. In total, there were 383 mathematics and science items included in the eighth-grade TIMSS assessment booklets. More detail on the distribution of new and trend items is included in table A3.

Table A3. Distribution of new and trend mathematics and science items in the TIMSS grade 4 and 8 assessments, by type: 2003

Response type	Grade 4			Grade 8		
	Total	New items	Trend items	Total	New items	Trend items
Total	313	243	70	383	230	153
Multiple choice	183	115	68	237	125	112
Constructed response	130	128	2	146	105	41
Mathematics	161	124	37	194	115	79
Multiple choice	92	55	37	128	69	59
Constructed response	69	69	0	66	46	20
Science	152	119	33	189	115	74
Multiple choice	91	60	31	109	56	53
Constructed response	61	59	2	80	59	21

SOURCE: Martin, M.O., Mullis, I.V.S., and Chrostowski, S.J. (2004). *TIMSS 2003 Technical Report: Findings from IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

The TIMSS 2003 frameworks included specifications for what are termed “problem-solving and inquiry” (PSI) tasks. PSI tasks were developed to assess how well students could draw on and integrate information and processes in mathematics and science as part of an investigation or in order to solve problems. The PSI tasks developed for TIMSS 2003 needed to be self-contained, involve minimal equipment, and be integrated into the main assessment without any special accommodations or additional testing time. While the PSI tasks are not full scientific investigations, the tasks were designed to require a basic understanding of the nature of science and mathematics, and to elicit some of the skills essential to the inquiry process. The tasks were designed to draw on students’ understandings of and abilities with formulating questions and hypotheses; designing investigations; collecting, representing, analyzing, and interpreting data; and drawing conclusions and developing explanations based on evidence.

The PSI tasks were assembled as longer blocks or clusters of items that, together, related to an overall theme (e.g., speciation). Nine PSI blocks were field-tested at fourth grade. Of the nine blocks, six blocks were eventually incorporated into the fourth-grade assessment. The six blocks covered both mathematics and science, focusing on geometry, measurement, number, life science, earth science, and physical science.

At eighth grade, 10 PSI blocks were field-tested. Of the 10 blocks, 7 blocks were eventually incorporated into the eighth-grade assessment. The seven blocks covered both mathematics and science, focusing on algebra, data, geometry, measurement, number, chemistry, physics, and life science. The PSI tasks were incorporated into the overall assessments and, thus, not reported separately at either grade level.

Design of Instruments

TIMSS 2003 included booklets containing assessment items as well as questionnaires submitted to principals, teachers, and students for response. The assessment booklets were constructed such that not all of the students responded to all of the items. This is consistent with other large-scale assessments, such as the National Assessment of Educational Progress. To keep the testing burden to a minimum, and to ensure broad subject-matter coverage, TIMSS used a rotated block design that included both mathematics and science items. That is, students encountered both mathematics and science items during the assessment. The 2003 fourth-grade assessment consisted of 12 booklets, each requiring approximately 72 minutes of response time. The 12 booklets were rotated among students, with each participating student completing 1 booklet only. The mathematics and science items were assembled into 14 blocks or clusters of items. Each block contained either mathematics items or science items only. The secure or trend items were included in 3 blocks, with the other

11 blocks containing replacement items. Each of the 12 booklets contained 6 blocks (in total).

The 2003 eighth-grade assessment also consisted of 12 booklets, each requiring approximately 90 minutes of response time. The 12 booklets were rotated among students, with each participating student completing 1 booklet only. The mathematics and science items were assembled into 14 blocks or clusters of items. Each block contained either mathematics items or science items only. The secure or trend items were included in 3 blocks, with the other 11 blocks containing replacement items. Each of the 12 booklets contained 6 blocks (in total).

As part of the design process, it was necessary to ensure that the booklets showed a distribution across the mathematics and science content domains as specified in the frameworks. The number of mathematics and science items in the fourth and eighth-grade TIMSS 2003 assessments is shown in table A4.

Table A4. Number of mathematics and science items in the TIMSS grade 4 and 8 assessments, by type and content domain: 2003

Content domain	Grade 4			Grade 8		
	Total	Response type		Total	Response type	
		Multiple choice	Constructed response		Multiple choice	Constructed response
Total items	313	183	130	383	237	146
Mathematics - Total	161	92	69	194	128	66
Number	63	30	33	57	43	14
Patterns, equations, and relationships	24	16	8	47	29	18
Measurement	33	23	10	31	19	12
Geometry	24	12	12	31	22	9
Data	17	11	6	28	15	13
Science - Total	152	91	61	189	109	80
Life science	65	41	24	54	29	25
Physical science	53	29	24	†	†	†
Earth science	34	21	13	31	22	9
Environmental science	†	†	†	27	10	17
Chemistry	†	†	†	31	20	11
Physics	†	†	†	46	28	18

†Not applicable. Content domain does not apply for the grade shown.

SOURCE: Martin, M.O., Mullis, I.V.S. and Chrostowski, S.J. (2004). *TIMSS 2003 Technical Report: Findings from IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Exhibit 2.21. Chestnut Hill, MA: Boston College.

In addition to the assessment booklets, TIMSS 2003 included questionnaires for principals, teachers, and students. As with prior iterations of TIMSS, the questionnaires used in TIMSS 2003 are based on prior versions of the questionnaires. The questionnaires were reviewed extensively by the national research coordinators from the participating countries as well as a Questionnaire Item Review Committee. Like the assessment booklets, all questionnaire items were field tested, and the results reviewed carefully. As a result, some of the questionnaire items needed to be revised prior to their inclusion in the final questionnaires. The questionnaires requested information to help provide a context for the performance scores, focusing on such topics as students' attitudes and beliefs about learning, student habits and homework, and their lives both in and outside of school; teachers' attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, instructional practices, and participation in professional development activities; and principals' viewpoints on policy and budget responsibilities, curriculum and instruction issues, student behavior, as well as descriptions of the organization of schools and courses.

Calculator Usage

Calculators were not permitted during the TIMSS fourth-grade assessment. However, the TIMSS policy on calculator use at the eighth grade was to give students the best opportunity to operate in settings that mirrored their classroom experiences. Beginning with 2003, calculators were permitted but not required for newly developed eighth-grade assessment materials. Participating countries could decide whether or not their students were allowed to use calculators for the new items; the United States allowed students to use calculators. Since calculators were not permitted at the eighth grade in the 1995 or 1999 assessments, the 2003 eighth-grade test booklets were designed so that trend items from these assessments were placed in the first half and new items in 2003 placed in the second half. Where countries chose to permit eighth-grade students to use calculators, they could use them for the second half of the booklet only.

Translation

Source versions of all instruments (assessment booklets, questionnaires and manuals) were prepared in English and translated into the primary language or languages of instruction in each country. In addition, it was sometimes necessary to adapt the instrument for cultural purposes, even in countries that use English as the primary language of instruction. All adaptations were reviewed and approved by the International Study Center to ensure they did not change the substance or intent of the question or answer choices. For example, proper names were sometimes changed to names that would be more familiar to students (e.g., Marja-leena to Maria).

Each country prepared translations of the instruments according to translation guidelines established by the International Study Center. Adaptations to the instruments were documented by each country, and submitted for review. The goal of the translation guidelines was to produce translated instruments of the highest quality that would provide comparable data across countries.

Translated instruments were verified by an independent, professional translation agency prior to final approval and printing of the instruments. Countries were required to submit copies of the final printed instruments to the International Study Center. Further details on the translation process can be found in the TIMSS 2003 Technical Report (Martin, Mullis, and Chrostowski 2004).

Test Administration and Quality Assurance

TIMSS 2003 emphasized the use of standardized procedures in all countries. Each country collected its own data, based on comprehensive manuals and trainings provided by the international project team to explain the survey's implementation, including precise instructions for the work of school coordinators and scripts for test administrators for use in testing sessions. Test administration in the United States was carried out by professional staff trained according to the international guidelines. School staff were asked only to assist with listings of students, identifying space for testing in the school, and specifying any parental consent procedures needed for sampled students.

Each country was responsible for conducting quality control procedures and describing this effort in the national research coordinators' report documenting procedures used in the study. In addition, the International Study Center considered it essential to monitor compliance with the standardized procedures. National research coordinators were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their countries. The International Study Center developed manuals for the monitors and briefed them in 2-day training sessions about TIMSS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities.

The national research coordinator in each country was responsible for scoring and coding of data in that country, following established guidelines. The national research coordinator and, sometimes, additional staff, attended scoring training sessions held by the International Study Center. The training sessions focused on the scoring rubrics and coding system employed in TIMSS. Participants were provided extensive practice in scoring example items over several days. Information on within-country agreement among coders was collected and documented by the International Study Center. Information on scoring and coding reliability was also used to calculate cross-country agreement among coders. Scoring reliability for TIMSS 2003 is provided in table A5.

Scoring Reliability

The TIMSS assessment items included both multiple choice and constructed-response items. A scoring rubric (guide) was created for every item included in the TIMSS assessments. These were carefully written and reviewed by national research coordinators and other experts as part of the field test of items, and revised accordingly.

Table A5. Within-country constructed-response scoring reliability for TIMSS grade 4 and 8 mathematics and science items, by exact percent score agreement and country: 2003

Country	Grade 4					
	Mathematics			Science		
	Average across items	Range		Average across items	Range	
		Min	Max		Min	Max
International average	99	92	100	96	85	100
Armenia	99	98	100	99	97	100
Australia	100	98	100	99	94	100
Belgium-Flemish	100	96	100	99	89	100
Chinese Taipei	99	83	100	98	89	100
Cyprus	98	91	100	94	76	100
England	99	91	100	98	87	100
Hong Kong SAR ¹	100	98	100	99	97	100
Hungary	98	91	100	95	80	100
Iran, Islamic Republic of	100	98	100	96	85	100
Italy	98	92	100	94	77	100
Japan	99	95	100	97	86	100
Latvia	98	87	100	96	82	100
Lithuania	97	77	100	93	81	100
Moldova, Republic of	100	100	100	100	100	100
Morocco	98	93	100	97	93	100
Netherlands	97	86	100	91	71	99
New Zealand	99	94	100	97	86	100
Norway	99	95	100	97	85	100
Philippines	99	96	100	97	89	100
Russian Federation	100	97	100	99	98	100
Scotland	99	98	100	98	90	100
Singapore	100	99	100	100	99	100
Slovenia	98	84	100	91	74	100
Tunisia	97	89	100	93	79	100
United States	97	88	100	93	70	100

See notes at end of table.

Table A5. Within-country constructed-response scoring reliability for TIMSS grade 4 and 8 mathematics and science items, by exact percent score agreement and country: 2003—Continued

Country	Grade 8					
	Mathematics			Science		
	Average across items	Range		Average across items	Range	
		Min	Max		Min	Max
International average	99	92	100	97	88	100
Armenia	99	94	100	98	92	100
Australia	100	97	100	99	94	100
Bahrain	99	98	100	98	94	100
Belgium-Flemish	99	96	100	97	89	100
Botswana	99	91	100	95	74	100
Bulgaria	96	70	100	91	72	99
Chile	99	95	100	97	91	100
Chinese Taipei	100	91	100	99	97	100
Cyprus	98	86	100	96	87	100
Egypt	100	97	100	100	98	100
Estonia	100	98	100	99	97	100
Ghana	99	97	100	98	93	100
Hong Kong SAR ¹	100	98	100	99	97	100
Hungary	98	90	100	96	87	100
Indonesia	98	90	100	96	87	100
Iran, Islamic Republic of	99	94	100	98	87	100
Israel	98	93	100	95	89	100
Italy	99	95	100	98	91	100
Japan	99	94	100	97	81	100
Jordan	99	98	100	99	97	100
Korea, Republic of	99	87	100	98	84	100
Latvia	98	90	100	94	78	100
Lebanon	100	94	100	100	98	100
Lithuania	97	71	100	90	69	100

See notes at end of table.

Table A5. Within-country constructed-response scoring reliability for TIMSS grade 4 and 8 mathematics and science items, by exact percent score agreement and country: 2003—Continued

Country	Grade 8					
	Mathematics			Science		
	Average across items	Range		Average across items	Range	
		Min	Max		Min	Max
Macedonia, Republic of	100	97	100	99	96	100
Malaysia	100	98	100	99	98	100
Moldova, Republic of	100	99	100	100	99	100
Morocco	97	89	100	94	86	100
Netherlands	97	84	100	90	70	100
New Zealand	99	96	100	98	92	100
Norway	98	91	100	95	83	100
Palestinian National Authority	99	94	100	95	82	100
Philippines	99	97	100	98	89	100
Romania	100	98	100	99	96	100
Russian Federation	99	95	100	99	92	100
Saudi Arabia	99	94	100	97	87	100
Scotland	99	95	100	97	89	100
Serbia	99	96	100	99	94	100
Singapore	100	98	100	100	99	100
Slovak Republic	100	98	100	99	95	100
Slovenia	97	86	100	90	70	100
South Africa	99	95	100	99	94	100
Sweden	98	89	100	92	76	100
Tunisia	98	89	100	98	90	100
United States	97	86	100	92	72	100

¹Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.

NOTE: To gather and document within-country agreement among scorers, systematic subsamples of at least 100 students' responses to each constructed-response item was coded independently by two readers. The agreement score indicates the degree of agreement among coders on marking student responses in the same way. See Mullis et al. (2004) and Martin et al. (2004) for more details.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2003.

Data Entry and Cleaning

Responsibility for data entry was taken by the national research coordinator from each country. The data collected for TIMSS 2003 were entered into data files with a common international format, as specified in the Manual for Entering the TIMSS 2003 Data. Data entry was facilitated by the use of a common software available to all participating countries (WinDEM). The software facilitated the checking and correction of data by providing various data consistency checks. The data were then sent to the IEA Data Processing Center (DPC) in Hamburg, Germany for cleaning. The DPC checked that the international data structure was followed; checked the identification system within and between files; corrected single case problems manually; and applied standard cleaning procedures to questionnaire files. Results of the data cleaning process were documented by the DPC. This documentation was then shared with the national research coordinator with specific questions to be addressed. The national research coordinator then provided the DPC with revisions to coding or solutions for anomalies. The DPC then compiled background univariate statistics and preliminary classical and Rasch Item Analysis. Detailed information on the entire data entry and cleaning process can be found in the TIMSS 2003 Technical Report (Martin, Mullis, and Chrostowski 2004).

Weighting, Scaling, and Plausible Values

Before the data were analyzed, responses from the groups of students assessed were assigned sampling weights to ensure that their representation in TIMSS 2003 results matched their actual percentage of the school population in the grade assessed. Based on these sampling weights, the analyses of TIMSS 2003 data were conducted in two major phases—scaling and estimation. During the scaling phase, item response theory (IRT) procedures were used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling were used to produce estimates of student achievement. Subsequent analyses related these achievement results to the background variables collected by TIMSS 2003.

Weighting

Responses from the groups of students were assigned sampling weights to adjust for over-representation or under-representation from a particular group. The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimators. The weight assigned to a student's responses is the inverse of the probability that the student would be selected for the sample. When responses are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual student assessed. Weighting also adjusts for various situations such as school and student nonresponse because data cannot be assumed to be randomly missing. The internationally defined weighting specifications for TIMSS require that each assessed student's sampling weight should be the product of (1) the inverse of the school's probability of selection, (2) an adjustment for school-level nonresponse, (3) the inverse of the classroom's probability of selection, and (4) an adjustment for student-level nonresponse. All TIMSS 1995, 1999 and 2003 analyses are conducted using sampling weights.

Scaling

TIMSS 1995, 1999, and 2003 used item response theory (IRT) methods to produce score scales that summarized the achievement results. With this method, the performance of a sample of students in a subject area or sub-area could be summarized on a single scale or a series of scales, even when different students had been administered different items. Because of the reporting requirements for TIMSS and because of the large number of background variables associated with the assessment, a large number of analyses had to be conducted. The procedures TIMSS used for the analyses were developed to produce accurate results for groups of students while limiting the testing burden on individual students. Furthermore, these procedures provided data that could be readily used in secondary analyses. IRT scaling provides estimates of item parameters (e.g., difficulty, discrimination) that define the relationship between the item and the underlying variable measured by the test. Parameters of the IRT model are estimated for each test question, with an overall scale being established as well as scales for each prede-

defined content area specified in the assessment framework. For example, the TIMSS 2003 eighth-grade assessment had five scales describing mathematics content strands, and science had scales for five fields of science.

TIMSS 1995 utilized a one parameter IRT model to produce score scales that summarized the achievement results. The TIMSS 1995 data were rescaled using a three-parameter IRT model to match the procedures used to scale the 1999 and 2003 TIMSS data. The three-parameter model was preferred to the one-parameter model because it can more accurately account for the differences among items in their ability to discriminate between students of high and low ability. After careful study of the rescaling process, the International Study Center concluded that the fit between the original TIMSS data and the rescaled TIMSS data met acceptable standards. However, as a result of rescaling, the average achievement scores of some countries changed from those initially reported in 1996 and 1997 (Peak 1996; NCES 1997). The rescaled TIMSS scores are included in this report.

Plausible Values

During the scaling phase, plausible values were used to characterize scale scores for students participating in the assessment. To keep student burden to a minimum, TIMSS administered a limited number of assessment items to each student—too few to produce accurate content-related scale scores for each student. To account for this, for each student, TIMSS generated five possible content-related scale scores that represented selections from the distribution of content-related scale scores of students with similar backgrounds who answered the assessment items the same way. The plausible-values technology is one way to ensure that the estimates of the average performance of student populations and the estimates of variability in those estimates are more accurate than those determined through traditional procedures, which estimate a single score for each student.

During the construction of plausible values, careful quality control steps ensured that the subpopulation estimates based on these plausible values were accurate. Plausible values were constructed separately for each national sample. TIMSS uses the plausible-val-

ues methodology to represent what the true performance of an individual might have been, had it been observed. This is done by using a small number of random draws from an empirically derived distribution of score values based on the student's observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. The draws from the distribution are different from one another to quantify the degree of precision (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performances. The TIMSS plausible values function like point estimates of scale scores for many purposes, but they are unlike true point estimates in several respects. They differ from one another for any particular student, and the amount of difference quantifies the spread in the underlying distribution of possible scale scores for that student. Because of the plausible-values approach, secondary researchers can use the TIMSS data to carry out a wide range of analyses.

Data Limitations

As with any study, there are limitations to TIMSS 2003 that researchers should take into consideration. Estimates produced using data from TIMSS 2003 are subject to two types of error, nonsampling and sampling errors. Nonsampling errors can be due to errors made in collecting and processing data. Sampling errors can occur because the data were collected from a sample rather than a complete census of the populations.

Nonsampling Errors

Nonsampling error is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. The sources of nonsampling errors are typically problems like unit and item nonresponse, the difference in respondents' interpretations of the meaning of the questions, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

Missing Data

There are four kinds of missing data: nonresponse, missing or invalid, not applicable, and not reached. *Nonresponse* data occurs when a respondent was expected to answer an item but no response was given. Responses that are *missing or invalid* occur in multiple-choice items where an invalid response is given. The code is not used for opened-ended questions. An item is *not applicable* when it is not possible for the respondent to answer the question. Finally, items that are *not reached* are consecutive missing values starting from the end of each test session. All four kinds of missing data are coded differently in the TIMSS 2003 database.

Missing background data are not included in the analyses for this report and are not imputed. In general, item response rates for variables discussed in this report were over the NCES standard of 85 percent to report without notation (table A6).

In general, it is difficult to identify and estimate either the amount of nonsampling error or the bias caused by this error. In TIMSS 2003, efforts were made to prevent such errors from occurring and to compensate for them when possible. For example, the design phase entailed a field test that evaluated items as well as the implementation procedures for the survey. It should also be recognized that most background information was obtained from students' self-reports, which are subject to respondent bias. One potential source of respondent bias in this survey was social desirability bias, for example, if students reported that they enjoyed mathematics.

Sampling Errors

Sampling errors occur when the discrepancy between a population characteristic and the sample estimate arises because not all members of the reference population are sampled for the survey. The size of the sample relative to the population and the variability of the population characteristics both influence the magnitude of sampling error. The particular sample of students in fourth and eighth grade from the 2002-03 school year was just one of many possible samples that could have been selected. Therefore, estimates produced from the TIMSS sample may differ from estimates that would have been produced had another student sample been drawn. This type of variability is called sampling error because it arises from using a sample of students in fourth or eighth grade, rather than all students in the grade in that year.

The standard error is a measure of the variability due to sampling when estimating a statistic. The approach used for calculating sampling variances in TIMSS was the Jackknife Repeated Replication (JRR). Standard errors can be used as a measure for the precision expected from a particular sample. Standard errors for all of the estimates are included in appendix C. The standard errors can be used to produce confidence intervals. There is a 95 percent chance that the true average lies within the range of 1.96 times the standard errors above or below the estimated score. For example, the average mathematics score for the U.S. eighth-grade students was 504 in 2003, and this statistic had a standard error of 3.3. Therefore, it can be stated with 95 percent confidence that the actual

Table A6. Weighted response rates for unimputed variables for TIMSS grade 4 and 8: 2003

Variable	Variable ID	Source of information	Grade 4		Grade 8	
			U.S. response rate	Range of response rates in other countries	U.S. response rate	Range of response rates in other countries
Sex	ITSEX	Classroom Tracking Form	100	94 – 100	100	92 – 100
Race/ethnicity	STRACE	Student Questionnaire	98	—	98	—
Free or reduced-priced lunch ¹	FRLUNCH	School Questionnaire	85	—	82	—

—Not available.

¹The response rate is calculated for public schools only.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2003.

average of U.S. eighth-grade students in 2003 was between 498 and 511 ($1.96 \times 3.3 = 6.5$; confidence interval = 504 ± 6.5).

Description of Background Variables

The international version of the TIMSS 2003 student, teacher and school questionnaires are available at <http://timss.bc.edu>. The U.S. versions of these questionnaires are available at <http://nces.ed.gov/timss>.

Race/Ethnicity

Students' race/ethnicity was obtained through student responses to a two-part question. Students were asked first whether they were Hispanic or Latino, and then asked whether they were members of the following racial groups: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, or White. Multiple responses to the race classification question were allowed. Results are shown separately for Asians, Blacks, Hispanics, and Whites. Students identifying themselves as Hispanic and also other races were included in the Hispanic group.

Poverty Level in Public Schools (Percentage of Students Eligible for Free or Reduced-price Lunch)

The poverty level in public schools was obtained from principal responses to the school questionnaire. The question asked what percentage of students at the school was eligible to receive free or reduced-price lunch through the National School Lunch Program around the first of October, 2002. The answers were grouped into five categories: less than 10 percent; 10 to 24.9 percent; 25 to 49.9 percent; 50 to 74.9 percent; and 75 percent or more. Analysis was limited to public schools only.

Confidentiality and Disclosure Limitations

The TIMSS 2003 data are hierarchical and include school data and student data from the participating schools. Confidentiality analyses for the United States were designed to provide reasonable assurance that public use data files issued by the IEA would not

allow identification of individual U.S. schools or students when compared against public data collections. Disclosure limitation included the identification and masking of potential disclosure-risk TIMSS schools and adding an additional measure of uncertainty of school, teacher, and student identification through random swapping of data elements within the student, teacher, and school files.

Statistical Procedures

Tests of Significance

Comparisons made in the text of this report have been tested for statistical significance. For example, in the commonly made comparison of country averages against the average of the United States, tests of statistical significance were used to establish whether or not the observed differences from the U.S. average were statistically significant. The estimation of the standard errors that are required in order to undertake the tests of significance is complicated by the complex sample and assessment designs which both generate error variance. Together they mandate a set of statistically complex procedures in order to estimate the correct standard errors. As a consequence, the estimated standard errors contain a sampling variance component estimated by Jackknife Repeated Replication (JRR); and, where the assessments are concerned, an additional imputation variance component arising from the assessment design. Details on the procedures used can be found in the WesVar 4.0 User's Guide (Westat 2000).

In almost all instances, the tests for significance used were standard t tests. These fell into two categories according to the nature of the comparison being made: comparisons of independent and non-independent samples. Before describing the t tests used, some background on the two types of comparisons is provided below:

The variance of a difference is equal to the sum of the variances of the two initial variables minus two times the covariance between the two initial variables. A sampling distribution has the same characteristics as any distribution, except that units consist of sample estimates and not observations. Therefore,

$$\sigma^2(\hat{\mu}_x - \hat{\mu}_y) = \sigma^2(\hat{\mu}_x) + \sigma^2(\hat{\mu}_y) - 2\text{cov}(\hat{\mu}_x, \hat{\mu}_y)$$

The sampling variance of a difference is equal to the sum of the two initial sampling variances minus two times the covariance between the two sampling distributions on the estimates.

If one wants to determine whether the girls' performance differs from the boys' performance, for example, then as for all statistical analyses, a null hypothesis has to be tested. In this particular example, it consists of computing the difference between the boys' performance mean and the girls' performance mean (or the inverse). The null hypothesis is:

$$H_0 : \hat{\mu}_{(boys)} - \hat{\mu}_{(girls)} = 0$$

To test this null hypothesis, the standard error on this difference is computed and then compared to the observed difference. The respective standard errors on the mean estimate for boys and girls ($\sigma(\hat{\mu}_{boys})$, $\sigma(\hat{\mu}_{girls})$) can be easily computed.

The expected value of the covariance will be equal to 0 if the two sampled groups are independent. If the two groups are not independent, as is the case with girls and boys attending the same schools within a country, or comparing a country mean with the international mean which includes that particular country, then the expected value of the covariance might differ from 0.

In TIMSS, country samples are independent. Therefore, for any comparison between two countries, the expected value of the covariance will be equal to 0, and thus the standard error on the estimate is:

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma_{(\hat{\theta}_i)}^2 + \sigma_{(\hat{\theta}_j)}^2}$$

with θ being any statistic.

Within a particular country, any sub-samples will be considered as independent only if the categorical variable used to define the sub-samples was used as an explicit stratification variable.

If sampled groups are not independent, the estimation of the covariance between, for instance, $\hat{\mu}_{(boys)}$

and $\hat{\mu}_{(girls)}$ would require the selection of several samples and then the analysis of the variation of $\hat{\mu}_{(boys)}$ in conjunction with $\hat{\mu}_{(girls)}$. Such a procedure is of course unrealistic. Therefore, as for any computation of a standard error in TIMSS, replication methods using the supplied replicate weights are used to estimate the standard error on a difference. Use of the replicate weights implicitly incorporates the covariance between the two estimates into the estimate of the standard error on the difference.

Thus, in simple comparisons of independent averages such as the U.S. average with other country averages, the following formula was used to compute the t statistic:

$$t = \frac{(est_1 - est_2)}{\sqrt{(se_1)^2 + (se_2)^2}}$$

Est_1 and est_2 are the estimates being compared (e.g., average of country A and the U.S. average) and se_1 and se_2 are the corresponding standard errors of these averages.

The second type of comparison used in this report occurred when comparing differences of non-subset, non-independent groups, such as when comparing the average scores of males versus females within the United States. In such comparisons, the following formula was used to compute the t statistic:

$$t = \frac{(est_{grp1} - est_{grp2})}{se(est_{grp1} - est_{grp2})}$$

Est_{grp1} and est_{grp2} are the non-independent group estimates being compared. $se(est_{grp1} - est_{grp2})$ is the standard error of the difference calculated using Jackknife Repeated Replication (JRR), which accounts for any covariance between the estimates for the two non-independent groups.

Effect size

Tests of statistical significance are, in part, influenced by sample sizes. To provide the reader with an increased understanding of the importance of the significant difference between student populations in the United States, effect sizes are included in the report. Effect sizes use standard deviations, rather than standard errors, and are therefore not influenced by the size of the student population samples. Following Cohen (1988) and Rosnow and Rosenthal (1996), effect size is calculated by finding the difference between the means of two groups and dividing that result by the pooled standard deviation of the two groups:

$$d = \frac{est_{grp1} - est_{grp2}}{sd_{pooled}}$$

Est_{grp1} and est_{grp2} are the student group estimates being compared. Sd_{pooled} is the pooled standard deviation of the groups being compared. The formula for the pooled standard deviation is as follows (Rosnow and Rosenthal 1996):

$$sd_{pooled} = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

Sd_1 and sd_2 are the standard deviations of the groups being compared. In social sciences, an effect size of .2 is considered small, one of .5 is of medium importance, and one of .8 or larger is considered large (Cohen 1988).

Country participation

Table A7 shows the countries that participated in TIMSS 2003 at fourth and eighth grades. The countries are grouped by continent. In addition, countries that are members of the Organization for Economic Cooperation and Development (OECD) are indicated with a check mark.

Table A7. Countries that participated in TIMSS grade 4 and 8 by continent and OECD membership: 2003

Grade 4		Grade 8	
Continent and country	OECD member	Continent and country	OECD member
Africa		Africa	
Morocco		Morocco	
Tunisia		Egypt	
		Ghana	
Asia		Tunisia	
Armenia		South Africa	
Chinese Taipei		Asia	
Hong Kong SAR ¹		Armenia	
Iran, Islamic Republic of		Bahrain	
Japan	3	Botswana	
Philippines		Bulgaria	
Singapore		Chinese Taipei	
Europe		Hong Kong SAR ¹	
Belgium-Flemish	3	Indonesia	
Cyprus		Iran, Islamic Republic of	
England	3	Israel	
Hungary	3	Japan	3
Italy	3	Jordan	
Latvia		Korea, Republic of	3
Lithuania		Lebanon	
Moldova, Republic of		Malaysia	
Netherlands	3	Palestinian National Authority	
Norway	3	Philippines	
Russian Federation		Saudi Arabia	
Scotland	3	Singapore	
Slovenia		Europe	
The Americas		Belgium-Flemish	3
United States	3	Cyprus	
Australia/Oceania		Estonia	
Australia	3	Hungary	3
New Zealand	3	Italy	3
		Latvia	
		Lithuania	
		Macedonia, Republic of	
		Moldova, Republic of	
		Netherlands	3
		Norway	3
		Romania	
		Russian Federation	
		Scotland	3
		Serbia	
		Slovak Republic	3
		Slovenia	
		Sweden	3
		The Americas	
		Chile	
		United States	3
		Australia/Oceania	
		Australia	3
		New Zealand	3

¹Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.

NOTE: The Organization for Economic Cooperation and Development (OECD) is an intergovernmental organization of 30 industrialized countries that serves as a forum for member countries to cooperate in research and policy development on social and economic topics of common interest.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study, 2003.
